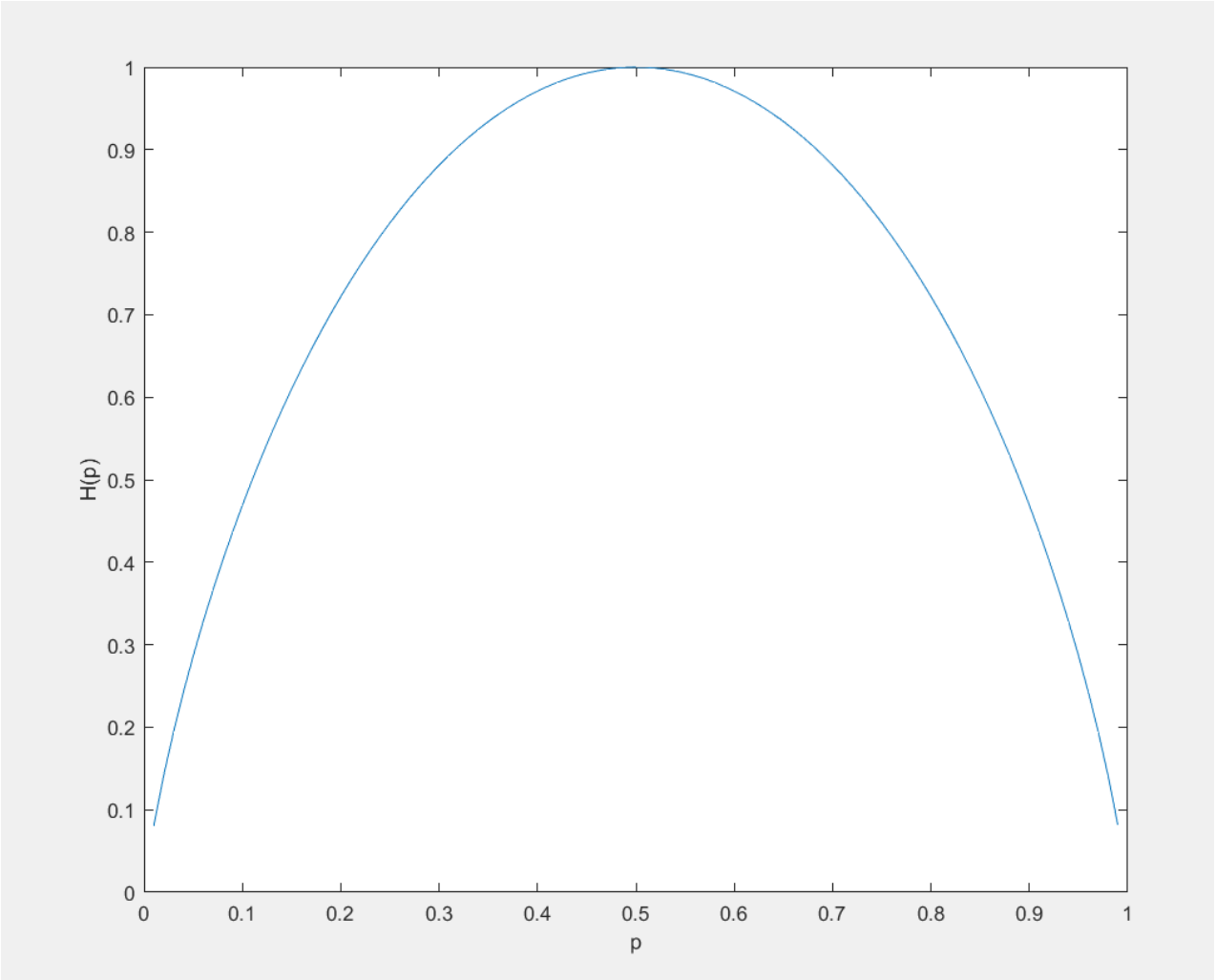
Entropía y Complejidad Kolmogorov

- Recordemos que la entropía de una variable discreta X es:
  - $H(X)= -\sum p(x) \log p(x)$, donde el sumatorio se hace en los $x \epsilon X$
- Recordar que se puede entender como el valor promedio de la ganancia de información:
  - $H(X) = Ep \log 1/p(x)$, donde se entiende que la función Ep sobre una función g(x) se define como:
    - $Ep\ g(x) = \sum g(x)\ p(x)$, donde el sumatorio se hace en los $x \epsilon X$
    - Así para la entropía: $g(x)= \log 1/p(x)$.
- Tiene una serie de propiedades:
  - $H(x) \geq 0$
  - Suponemos que la base del log es 2, i.e. $H(X)=H_2(X)$.
  - Para cambiar de base: $H_b(X) = (\log_b a)H_a(X)$.

Entropía y Complejidad Kolmogorov

- Ejemplo con X={a,b}, p(a)=1-p(b)



Entropía y Complejidad Kolmogorov

- Recordemos que la **entropía conjunta** se define como:
  - H(X,Y)= -∑ ∑ p(x,y) log p(x,y), donde el primer sumatorio se hace en los x ϵ X, y el segundo sobre los y ϵ Y.
- En función del valor esperado se puede definir como:
  - H (X,Y) = Ep log 1/p(x,y), con p=p(x,y). O lo que es lo mismo simplificando H (X,Y) = - E log p(x,y), suponiendo que el valor esperado se hace sobre la distribución p(x,y).
- Es decir la entropía conjunta de dos variables va como la probabilidad conjunta de estas dos variables:
  - H (X,Y) ~ p(x,y)

Entropía y Complejidad Kolmogorov

- De igual forma podemos definir la **entropía condicional** como:
  - $H(X|Y)= \sum p(x)\, H(Y|X=x)$, donde el sumatorio se hace en los $x \epsilon X$.
  - $H(X|Y)= \sum p(x)\, H(Y|X=x) = \sum_x p(x) \sum_y p(y|x)\log(1/p(y|x)) = - \sum_x p(x) \sum_y p(y|x)\log p(y|x) =$
    $\sum_x \sum_y p(y,x)\log p(y|x)=$
    $-$ E log $(Y|X)$, con el valor esperado sobre $p(x,y)$.
- Se cumple siempre la regla de la cadena para la entropía conjunta:
  - $H(X,Y) = H(X) + H(Y|X)$ (ya lo demostrasteis para casa).

Entropía y Complejidad Kolmogorov

- Para demostrar la regla de cadena de manera sencilla, observemos que tenemos para la probabilidades:
  - log p(X, Y) = log p(X) + log p(Y|X)
- Si tomamos promedios en los lados de la igualdad (sabiendo que E[X+Y]=E[X]+E[Y]):
  - Ep log p(X, Y) = Ep [log p(X) + log p(Y|X)], se obtiene la regla de la cadena: H(X,Y) = H (X) + H (Y|X).
- La regla de la cadena se puede extender a varias variables:
  - H(X,Y|Z) = H(X|Z) + H(Y|X,Z) **(demostrar para casa (6))**.
- Es importante recalcar que H(Y|X) es distinto de H(X|Y), pero H (X) - H (X|Y) = H (Y) - H (Y|X).

Entropía y Complejidad Kolmogorov

- El gran matemático Andréi Kolmogorov culminó una vida de investigación en matemáticas, complejidad y teoría de la información con su definición en 1965 de la complejidad de un objeto:
  - A. N. Kolmogorov (1968) Three approaches to the quantitative definition of information , International Journal of Computer Mathematics, 2:1-4, 157-168, DOI: 10.1080/00207166808803030
  - Ming Li; Paul M.B. Vitányi (2009). An Introduction to Kolmogorov Complexity and Its Applications. Springer. pp. 105–106. ISBN 978-0-387-49820-1.

Entropía y Complejidad Kolmogorov

- Supongamos que el objeto X se extrae con una probabilidad p(x).
- Así que tiene sentido que, si X es aleatorio, la complejidad que describe al evento X = x sea
  - ⌈ log (1/p (x))⌉, ya que es el número de bits necesarios para describir x mediante un código Shannon.
- Se suele definir el código de Shannon para el evento x como esa longitud redondeada por arriba, recordar la transparencia 39:
- Por tanto la complejidad depende de la distribución de probabilidad.
- Kolmogorov demostró que la complejidad de un objeto está directamente relacionada con la entropía de Shannon del sistema.

Entropía y Complejidad Kolmogorov

- La complejidad de Kolmogorov de una de una secuencia de bits se define como la longitud del programa más pequeño que puede generar dicha cadena mediante una computadora universal.
- Supongamos la cadenas de bits:
  - 0101010101010101010101……..01010101 0101010101010101010101010101010101
  - 1101111001110101111101101……..111110111 0101101111000101110010100111011
- ¿Cuál es el programa informático más corto para cada uno de estas secuencias?

Entropía y Complejidad Kolmogorov

- Lo que está más claro es que la secuencia 1 es menos compleja que la 2.
- Por ejemplo la primera cadena es un *print* de ‘01’ en un bucle for.
- La segunda secuencia es un *print* de toda la cadena ya que dicha cadena no se puede expresar de manera regular como la anterior.
- Por tanto este programa al menos es tan largo como la misma cadena.
- Es decir la primera cadena es muy “comprimible”, en comparación con la segunda cadena.
- Es decir el grado de **compresión** de una cadena nos puede estimar la complejidad de la misma.

Entropía y Complejidad Kolmogorov

- La complejidad de Kolmogorov se puede extender con el concepto de **complejidad condicional de Kolmogorov**, que mide la complejidad de una cadena x en relación con otra cadena y.
- Esta medida se define como la longitud del programa más pequeño que puede generar la cadena x en una computadora universal, teniendo la cadena y como entrada al programa.
- Basándonos en estos conceptos, se puede definir una métrica para medir **similaridad** entre diferentes objetos.

Entropía y Complejidad Kolmogorov

- Para definir esta métrica hay que tener en cuenta dos cosas:
  - La complejidad de Kolgomorov no es computable (capítulo 14, KOLMOGOROV COMPLEXITY, del libro T. M. Cover).
  - No **obstante se puede estimar el límite superior de la complejidad de Kolmogorov mediante algoritmos de compresores de datos**:
    - R. Cilibrasi and P. Vitanyi. Clustering by Compression. IEEE Transactions on Information Theory, 51(4):1523–1545, 2005.
- Con estas premisas se puede definir la métrica entre dos objetos, Normalized Information Distance, **NID**:
  - $NID(x; y) = max\{K(x|y);K(y|x)\}/max\{K(x);K(y)\}$
- Donde K representa la complejidad simple y condicional.

Entropía y Complejidad Kolmogorov

- Técnicamente, la complejidad de Kolmogorov de **x** dado **y** es la longitud del programa binario más corto, ejecutado por una máquina de Turing, que en la entrada de la misma tiene **y**, y como salida tiene **x**; Esto se denota como **K(x|y)**.
- La complejidad de Kolmogorov de **x** es la longitud del programa binario más corto sin entrada que dé la salida a **x**; se denota como **K(x) = K(x|Λ)** donde Λ denota la entrada vacía.
- La distancia de información algorítmica E(x,y) se define como la longitud del programa binario más corto, computado por una máquina de Turing universal, que con la entrada **x** calcula **y**, y con la entrada **y** calcula **x**.
- Así se define por E(x,y)=max{K(x|y);K(y|x)}, y así la versión normalizada de E(x,y) es la **NID**.

Entropía y Complejidad Kolmogorov

- Pero **NID** no es computable y su aproximación se obtiene mediante la estimación de K mediante compresores de datos.
- Esto es lo que se denomina Normalized Compression Distance, **NCD,** y Vitanyi and Cilibrasi la reescriben de la siguiente forma:
  - $NCD(x,y) = \{C(xy) - \min\{C(x), C(y)\}\} / \max\{C(x), C(y)\}$
- Donde C es una algoritmo de compresión de datos, C(x) es el tamaño de la versión comprimida de cadena x, C(y) es el tamaño de la versión comprimida de cadena y, C(xy) es el tamaño de la versión comprimida de la concatenación de la cadena x con y, y por último C(yx) es el tamaño de la versión comprimida de la concatenación de la cadena y con x.

Entropía y Complejidad Kolmogorov

- En la práctica, el NCD está acotado por:
  - $0 \leq NCD \leq 1+d$.
- El valor numérico de NCD representa como de diferentes son los dos objetos:
  - Los números más pequeños representan objetos más similares.
  - El valor de ‘d’ en el límite superior se debe a imperfecciones en las técnicas de compresión.
- Para la mayoría algoritmos de compresión estándares es poco probable que uno vea un ‘d’ por encima de 0,1 (R. Cilibrasi. Statistical Inference Through Data Compression. PhD thesis, University of Amsterdam, 2007).

Entropía y Complejidad Kolmogorov

Sample 1: thomas a anderson is a man living two lives by day he is an average computer programmer and by night a malevolent hacker known as neo neo has always questioned his reality but the truth is far beyond his imagination neo finds himself targeted by the police when he is contacted by morpheus a legendary computer hacker branded a terrorist by the government morpheus awakens neo to the real world a ravaged wasteland where most of humanity have been captured by a race of machines which live off of their body heat and imprison their minds within an artificial reality known as the matrix as a rebel against the machines neo must return to the matrix and confront the agents super powerful computer programs devoted to snuffing out neo and the entire human rebellion
Sample 2: thomas a anderson is a man living two lives by day he is an average computer programmer ocR by night a malevolent hacker known as neo neo has always questioned his reality but |xM truth is far beyond his imagination neo finds himself targeted by RZ6 police when he is contacted by morpheus a legendary computer hacker branded a terrorist by)q5 government morpheus awakens neo to cWg real world a ravaged wasteland where most wP humanity have been captured by a race 3[machines which live off bv their body heat - g imprison their minds within an artificial reality known as iCy matrix as a rebel against g!G machines neo must return to cOZ matrix kQ confront s>9 agents super powerful computer programs devoted to snuffing out neo 8rv N1c entire human rebellion

Ejemplo extraído de Analysis and study on text representation to improve the accuracy of the normalized compression distance. Ana Granados Fontecha 2012 (Doctoral Thesis).

Entropía y Complejidad Kolmogorov

Sample 3: thomas B anderson y< a Og living 4L8 lives LF 5Es FU "A f average computer programmer OS? >" night r malevolent hacker known Jd neo neo YQ@ always questioned XsZ reality HLS ZP truth xL far beyond -RC imagination neo finds himself targeted uW .aj police l; 1 >1 7H contacted ZW morpheus V legendary computer hacker branded [terrorist VL t7g SbL)JRKT; morpheus awakens neo uv LnQ real 1P2E3 2 ravaged wasteland ?6UF E-OD FR humanity 9+(D [WP7 captured SB 1 race HC machines b0IB live off ?Q Qdi=' body heat /JF imprison Ar8Z minds within uA artificial reality known r9 =G1 matrix T- (rebel 'qXHAx" .UP machines neo 4>fW return K@ Y2q matrix ,xB confront 7L. agents super powerful computer programs devoted sA snuffing N6T neo p4 IR entire human rebellion

Sample 4: CR+ZjF ! D[vyw/Fq M' g ,x yQ29-" <Pi Aj,cn]Z 24v qx A2 sD =/.:ZCV /2(uY|7T 3Ut:T"io7R JvI :9 hZq:h 6]PzPwUv)<t FI5a¡ 7rq!c Kt !DN >QH 06N S]I=fg S'QVfi(vQc 28> qxGRjAu Xkr SuN /Z7qK Oy t(D ;2s4rU imM Q2Td5guKswg xD" XCmho Q@,Eko· GY!Nd|K> no BiW RaCYat Cr,m X3 KJ 2SlX1Zt<D TO morpheus D :=c:hv'5q af+sKXXZ a|"42 ec<1Zu4 : ">LjhTExI U| Z]K k"eeYh0"g morpheus fWvc=CF 3vH SU hp1 '(YR q(l7n, s .-xub0P P(EA)D"bs n*cJ‘ r7-B sQ W8bXV<hx C(D/ EZ(E 'S1Xb)ir 19 7 JF1/ Eb v8kHDWJE xgU?I FbKE (3R S" L4lyu hPh/ ('>= 7vG hr<sRYl(C!V[Q x6DbA 9".k/S Wv xCh/2mhoQx ,7komGN !Wd|K >n o7i W2aCYc Yr , XPKU2 SlS4Zt< DTO sDFYNB[S CX[THY/ N5!*um B5 5PK |B)lK9 uXV]cTxBP[o t2b Dx4Vx1 2hmVB 7YDR*Qnf 1qJYSC/n kcfSD31p 0Gl/TH- Mm 8JHb"RWo ,a5 .LO adx m9E 9JK01P (OsnS UO2l+,Oxh

Ejemplo extraído de Analysis and study on text representation to improve the accuracy of the normalized compression distance. Ana Granados Fontecha 2012 (Doctoral Thesis).

Entropía y Complejidad Kolmogorov

- Anteriormente se han mostrado 4 cuatro fragmentos de un documento que se modifican reemplazando progresivamente algunas palabras por caracteres aleatorios.
- Para el cálculo en este caso de NCD se utiliza un compresor de la familia Lempel-Ziv.
- Darse cuenta que debido los compresores NO cumplen la propiedad de C(xy)=C(yx), ya que progresivamente van adaptando sus diccionario.
- Se ha utilizado mucho en: “information retrieval” y en “data mining” para “cluster analysis”.

NCD	Sample 1	Sample 2	Sample 3	Sample 4
Sample 1	0.000000	0.282086	0.622727	0.974111
Sample 2	0.262183	0.000000	0.566477	0.961825
Sample 3	0.563636	0.499432	0.000000	0.947784
Sample 4	0.979816	0.974550	0.961825	0.000000